D3Rings: A fast and accurate method for ring system identification and deep generation of drug-like cyclic compounds

Minfei Ma[1,2], Xinben Zhang[1], Liping Zhou[1,2], Zijian Han[1,2], Yulong Shi[1,2], Jintian Li[1,2], Leyun Wu[1,2], Zhijian Xu[1,2*], Weiliang Zhu[1,2*]

[1]Stake Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China

[2]School of Pharmacy, University of Chinese Academy of Sciences, Beijing 100049, China

[*]To whom correspondence should be addressed.

Phone: +86-21-50805020 (W.Z.), Fax: +86-21-50807088 (W.Z.), Email: zjxu@simm.ac.cn (Z.X.), wlzhu@simm.ac.cn (W.Z.).

ABSTRACT

Continuous exploration of the chemical space of molecules to find ligands with high affinity and specificity for specific targets is an important topic in drug discovery. A focus on cyclic compounds, particularly natural compounds with diverse scaffolds, provides important insights into novel molecular structures for drug design. However, the complexity of their ring structures has hindered the applicability of widely accepted methods and software for the systematic identification and classification of cyclic compounds. Herein, we successfully developed a new method, D3Rings, to identify acyclic, monocyclic, spiro ring, fused and bridged ring, and cage ring compounds as well as macrocyclic compounds. By using D3Rings, we completed the statistics of cyclic compounds in 3 different databases, e.g., ChEMBL, DrugBank, and COCONUT. The results demonstrated the richness of ring structures in natural products, especially spiro, macrocycles, fused and bridged rings. Based on this, three deep generative models, namely VAE, AAE, and CharRNN, were trained and used to construct two datasets similar to DrugBank and COCONUT but 10 times larger than them. The enlarged datasets were then used to explore the molecular chemical space, focusing on complex ring structures, for novel drug discovery and development. Docking experiments with the newly generated COCONUT-like dataset against three SARS-CoV-2 target proteins revealed that an expanded compound database improves molecular docking results. Cyclic structures were exhibited the best docking scores among the top-ranked docking molecules. These results suggest the importance of exploring the chemical space of structurally novel cyclic compounds and continuous expansion of the library of drug-like compounds to facilitate the discovery of potent ligands with high binding affinity to specific targets. D3Rings is now freely available at http://www.d3pharma.com/D3Rings/.

INTRODUCTION

Cyclic compounds play an important role in drug development. Among the top 200 pharmaceuticals based on the retail sales in 2021, compiled by M. Haziq Qureshi from the University of Arizona, 113 are small-molecule drugs and 112 are cyclic compounds. Among the 112 cyclic-compound drugs, 76, 8, and 2 drugs contain fused or bridged ring, macrocycle, and spiro ring structures, respectively. Cyclic compounds also account for a considerable proportion of recently approved small-molecule drugs against COVID-19, such as Paxlovid, a combination of the 3CLpro inhibitor Nirmatrelvir and the antiviral small molecule Ritonavir developed by Pfizer, the RdRp inhibitor Molnupiravir jointly developed by Merck and Ridgeback, the 3CLpro inhibitor Ensitrelvir developed by Shionogi, and the RdRp inhibitor VV116 developed in China.[1–5]

Studies on molecular skeleton analysis have been evolving due to advancement in computational chemistry, cheminformatics, etc. A conventional method to define the scaffold of a molecule is the Murcko framework proposed by Bemis and Murcko.[6,7] This method breaks down a molecule into ring systems, linkers, side chains, and the Murcko framework—a combination of ring systems and the linkers in the molecule. Based on the Murcko framework, an approach called Scaffold Tree (ST) introduces a hierarchical tree to describe ring systems, where rings are iteratively pruned one by one according to a set of prioritization rules until only one ring remains.[8] Similarly, the SCONP approach first trims all the terminal side chains in a molecule to obtain a scaffold.[9] Scaffolds obtained in this manner are hierarchically grouped by establishing parent–child relationships among the scaffolds, with parent scaffolds representing substructures of the child scaffolds. The parent–child relationships are then combined into a classification tree.

Graph theoretic approaches have played a pivotal role in classifying and analyzing of ring systems within molecular structures. By representing chemical compounds as graphs, with atoms as nodes and bonds as edges, graph theory provides a powerful framework for understanding the connectivity and topology of ring systems.[10] Graph-based techniques, such as the smallest set of smallest rings (SSSR) method and ring perception algorithms, have been developed to characterize and classify ring systems based on their topological properties, symmetry, and aromaticity.[10,11] Herein, a new strategy for analyzing cyclic compounds based on the SSSR algorithm, which identifies and classifies molecules based on their linkage to the ring is described.

Novel structures are always pursued by drug developers, necessitating artificially generated datasets with cyclic structures similar to marketed drugs or natural products for virtual screening and drug design. Traditional approaches to generating new molecular structures involve combining or recombining of chemical fragments, which are not feasible for highly complicated ring structures.

In recent years, with the development of big data in chemistry and the advancement of artificial intelligence, deep generative models have shown enormous potential in accelerating drug discovery and introducing innovation to exploration within the chemical space.[12–14] Compared to traditional molecular generation methods that rely on the combination or recombination of chemical fragments, deep generative models offer a new way that does not directly rely on structural similarity.[15] Therefore, utilizing deep generative models can expand the chemical space of drug-like compounds, increase the diversity of cyclic compound structures, and establish structurally novel and larger-scale compound libraries for future drug research.[16]

Herein, we focused on drug-like and natural product–like cyclic compounds. First, we developed D3Rings, a fast and accurate ring system identification and classification method. Second, using D3Rings, we systematically classified cyclic compounds across various databases. Third, employing three deep generative models—VAE, AAE, and CharRNN—we constructed large-scale datasets of natural product–like and drug-like molecules. Lastly, through the virtual screening of millions of compounds, we demonstrated that large-scale databases enhance the discovery of structurally diverse, high-affinity drug-like compounds in the context of drug–target interactions.

MATERIALS AND METHODS

Data collection

Our experiments utilized three molecular databases: ChEMBL30, a drug-like compound dataset; DrugBank (5.1.9), containing approved/investigational drug information; and COCONUT, an extensive collection of open natural product dataset, comprising the most complete up-to-date dataset of natural product compounds and natural product–like compounds.[17–19] These three databases have different characteristics owing to their respective data sources. We obtained molecules from the ChEMBL30, DrugBank (5.1.9), and COCONUT (January 2022) databases in SMILES string representation.[20] To ensure compatibility with the RDKit v.2020.09.1 toolkit in Python, we removed very few molecules that could not be processed.

The classification of molecules in molecular datasets

Figure 1A–C illustrates the classification of molecules according to the presence or absence of ring structures in the molecule and the characteristics of the ring linkage. Molecules are categorized into acyclic compounds (containing no ring structures), monocyclic compounds (containing only one ring or no direct linkage between rings), and joined ring compounds (two or more rings directly linked to each other). Especially, the joined ring compounds exhibit different ways that two or more rings can be connected (Figure 1C). If two or more rings are linked by one common atom with twisted structures, the structure is called a spiro ring structure.[21] If two or more rings share two or more atoms joined together by common ring edges, then the resulting structure is called a fused ring structure, where the rings share two adjacent atoms. Conversely, if the rings share two nonadjacent atoms with one or more atoms between them, then it is called a bridged ring structure. Among the fused and bridged ring structures, there is a special type of structure, generally a hollow cage with a three-dimensional structure, called a cage ring structure.[22] Additionally, we screened macrocyclic molecules, a common structural feature in cyclic compounds: a ring structure of 12 or more atoms (Figure 1D).[23]

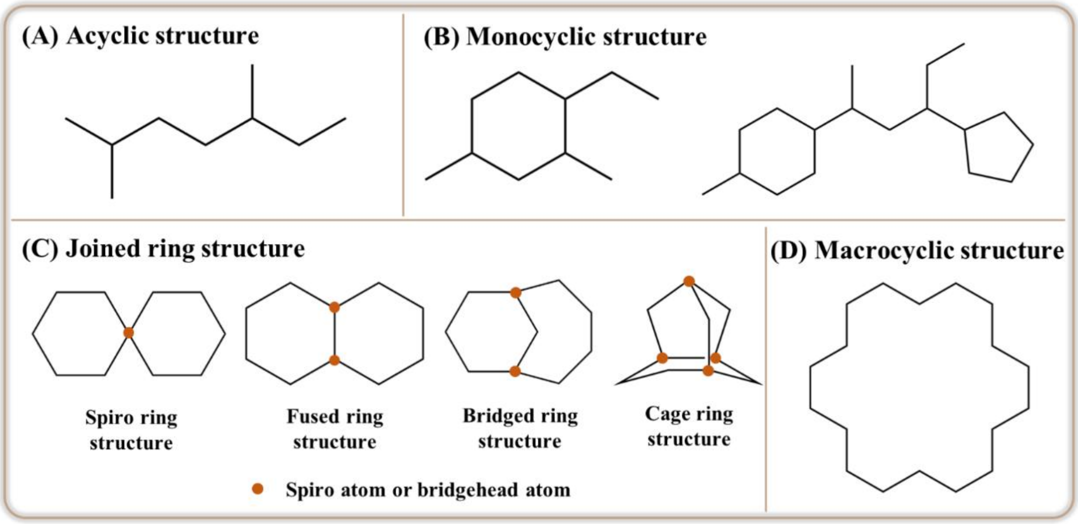

Figure 1. Schematic diagram of (A) acyclic structure, (B) monocyclic structure, (C) joined ring structure, and (D) macrocyclic structure.

Using our newly developed method D3Rings, we categorized molecules in ChEMBL30, DrugBank (5.1.9), and COCONUT (January 2022) into acyclic, monocyclic, spiro ring, fused and bridged ring, cage ring, and macrocyclic compounds based on their structural characteristics.

Ring structure classification method: D3Rings

The flow of D3Rings for molecular classification is shown in Figure 2A. The program first identifies the ring structure in the molecule based on the SSSR method, checking for directly connected molecules in the ring system. Subsequently, the program performs further identification based on the unique properties of each type of joined ring compound. For spiro ring compounds, we need to find molecules with a single bridgehead atom in their structure, while macrocyclic compounds are identified by locating molecules with a ring structure that contains 12 or more atoms. Especially, fused and bridged ring compounds are subdivided according to the maximum number of joined rings in the molecule. Figure 2B illustrates the calculation of the maximum number of joined rings in a molecule, achieved by breaking the molecule from the acyclic bond to obtain several molecular fragments. The number of joined rings in each molecular fragment is determined using the following formula: the number of joined rings = the number of ring bonds − the number of ring atoms + 1. The maximum value of joined rings among all fragments represents the final maximum number of joined rings in one molecule. Notably, for molecules containing multiple ring assemblies, we classify the molecules in terms of the maximum number of directly connected rings in the ring system. The entire process is executed using our Python program, mainly the RDKit v.2020.09.1 package. D3Rings is now available at http://www.d3pharma.com/D3Rings/.

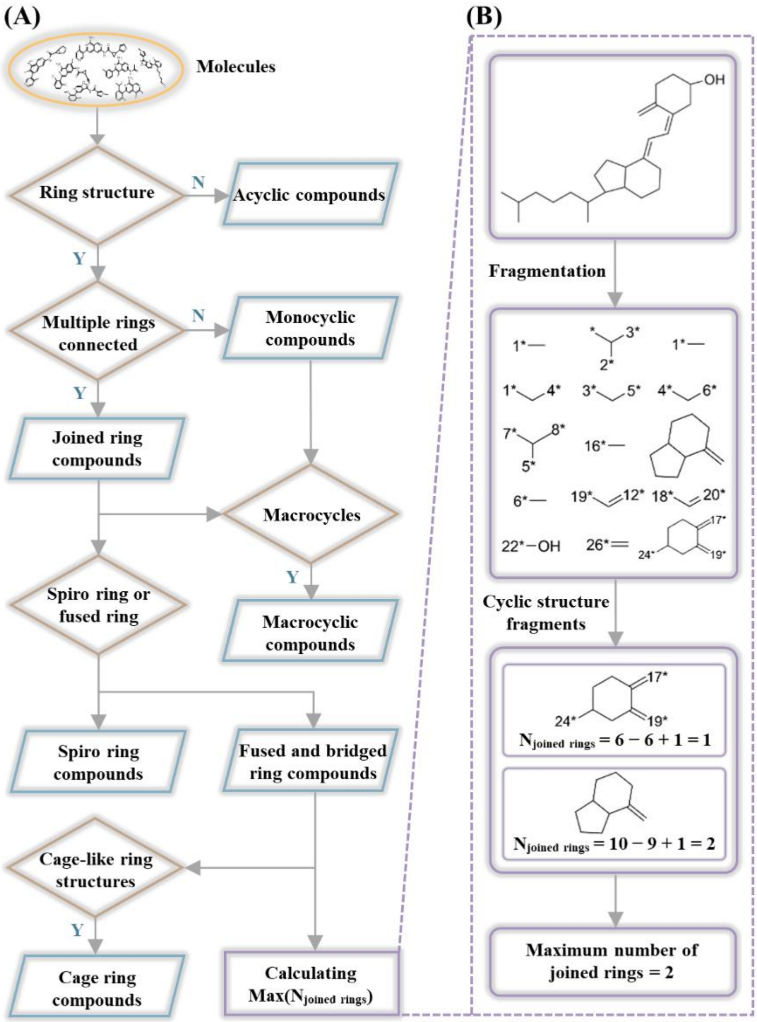

Figure 2. Strategies for D3Rings. (A) Flow chart of the molecular classification procedure. (B) Schematic diagram of the strategy for calculating the maximum number of joined rings.

Statistics on halogenated compounds in molecular datasets

To investigate the distribution of halogenated compounds across various databases, we selected all compounds, spiro ring compounds, fused and bridged ring compounds, and macrocyclic compounds from ChEMBL30, DrugBank (5.1.9), and COCONUT (January 2022), respectively, as sample pools. Subsequently, we counted the proportions of halogen-containing compounds, compounds with a single type of halogen, and the distribution of the number of halogen atoms within halogenated compounds in each sample pool.

Deep generative models for building new molecular databases

Deep generative models have emerged as new tools for exploring the molecular chemistry space. Herein, we used three deep generative models to build the molecular datasets: variational autoencoders (VAE), adversarial autoencoders (AAE), and character-level recurrent neural networks (CharRNN).

VAE operates by encoding high-dimensional input into a low-dimensional space (latent space) via an encoder and then decoding it back to high-dimensional output via a decoder (Figure 3A).[24,25] The loss function used by VAE for training includes two components: reconstruction loss and Kullback–Leibler divergence (KL divergence).[26,27] We used a gated recurrent unit (GRU) bidirectional recurrent neural network with a linear output layer as an encoder and a three-layer GRU with dropout layers as a decoder. VAE is designed to generate novel cyclic drug molecules with structural properties similar to the training set. VAE models can be trained on a dataset of cyclic drugs and used to explore the chemical space for new drugs.

AAE is also a probabilistic autoencoder like VAE but uses the generative adversarial networks to perform variational inference that replaces the KL divergence in VAE (Figure 3B).[28,29] The encoder in this work incorporates a single-layer long short-term memory (LSTM), and the decoder comprises a two-layer LSTM. A two-layer fully connected linear neural network with an exponential linear unit activation function forms the discriminator. AAE can be trained on cyclic drug datasets to generate molecules possessing desired properties and aligning well with the training data.

CharRNN is a deep learning model for modeling a series of characters; the distribution of the next character is predicted based on the observed characters.[30,31] We used three-layers LSTM cells with a dropout layer and Softmax as the output layer. CharRNN is trained by maximizing the log-likelihood of the training data (Figure 3C). CharRNN can be used to generate novel cyclic drug molecules by training on a large corpus of known compounds. RNN are particularly useful when the desired properties of cyclic drug molecules are closely related to their sequential characteristics, such as specific functional groups or substructures.

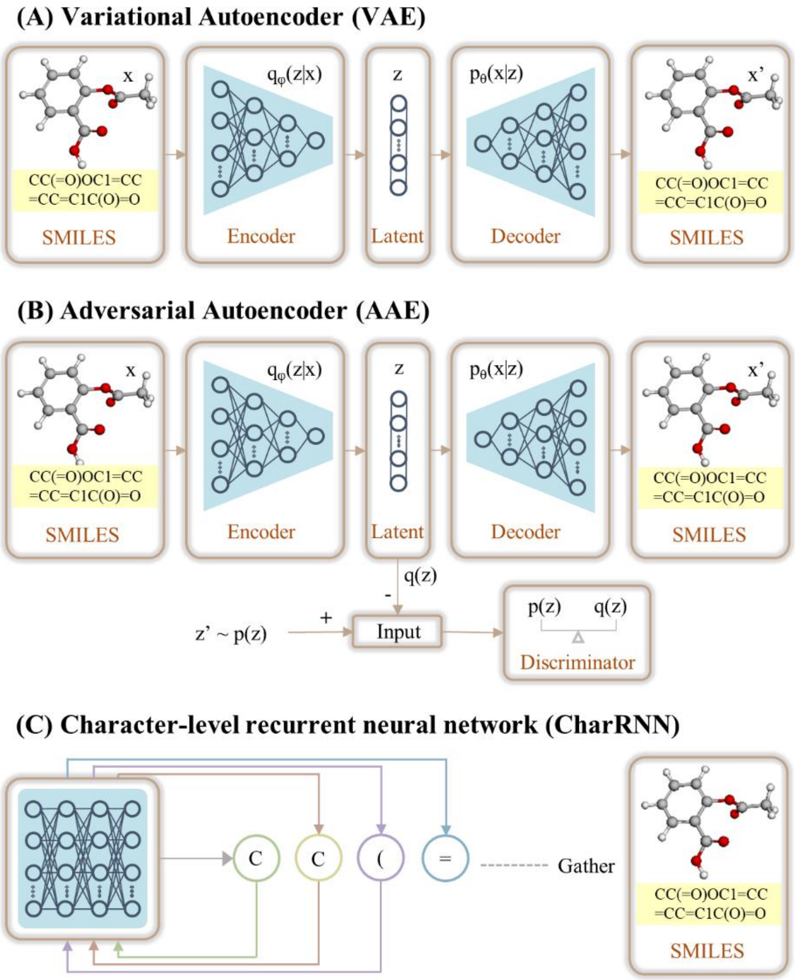
(A) Variational Autoencoder (VAE)
x
CC(=O)OC1=CC =CC=C1C(O)=O
SMILES
$q_\varphi(z|x)$
Encoder
z
Latent
$p_\theta(x|z)$
Decoder
x'
CC(=O)OC1=CC =CC=C1C(O)=O
SMILES
(B) Adversarial Autoencoder (AAE)
x
CC(=O)OC1=CC =CC=C1C(O)=O
SMILES
$q_\varphi(z|x)$
Encoder
z
Latent
$p_\theta(x|z)$
Decoder
x'
CC(=O)OC1=CC =CC=C1C(O)=O
SMILES
q(z)
-
+
z' ~ p(z)
Input
p(z)
q(z)
Discriminator
(C) Character-level recurrent neural network (CharRNN)
C
C
(
=
Gather
CC(=O)OC1=CC =CC=C1C(O)=O
SMILES

Figure 3. Frameworks of (A) variational autoencoder, (B) adversarial autoencoder, and (C) character-level recurrent neural network.

We split the DrugBank (5.1.9) and COCONUT (January 2022) databases into training and validation sets at a ratio of 9:1, respectively. The DrugBank training set had 10,154 molecules, while the COCONUT training set had 366,228 molecules. Correspondingly, the DrugBank validation set had 1,129 molecules, and the

COCONUT validation set had 40,692 molecules. Subsequently, VAE, AAE, and CharRNN were used to build DrugBank-like (4,185,929 molecules) and COCONUT-like (119,381 molecules) databases, each expanding the original database by approximately tenfold. In this process, SMILES was used as input and output representations.

Evaluation and statistics for generating molecular databases

To evaluate the ability of the models to generate molecules and test the performance of these molecules in the newly generated drug-like and natural product–like database, we proposed a set of metrics in Molecular Sets (MOSES) to assess the generated molecules.[32] These metrics include molecular validity (Valid), molecular novelty (Novelty), internal diversity (Internal diversity), uniqueness (Unique), fragment similarity (Frag), scaffold similarity (Scaff), and similarity to nearest neighbor (SNN). Validity, Novelty, Internal diversity, and Unique were calculated for the difference in intact molecules between training set and generated set. The other three metrics compare the differences between the BRICS fragments (Frag), Bemis–Murcko scaffolds (Scaff), and molecular fingerprints (SNN) obtained using the training and generated sets, respectively. For the uniqueness metric, we calculated Unique 10k in our experiments, representing the uniqueness of the top 10,000 valid molecules in the generated compound dataset. In addition, we calculated the distributions of molecular weight (MW), oil–water partition coefficient (LogP), quantitative estimation of drug-likeness (QED), and synthetic accessibility (SA) for the generated and training molecule sets using MOSES.

To verify the molecular composition of the new database is consistent with the original database, the same statistics for acyclic, monocyclic, spiro ring, fused and bridged ring, macrocyclic, and halogenated compounds were performed for the DrugBank- and COCONUT-like databases using D3Rings.

Exploring the impact of molecular dataset size on virtual screening results

Virtual screening via molecular docking is useful for discovering active compounds from a chemical library.[33,34,35] To explore the necessity of generating large-scale chemical datasets for virtual screening, molecular docking was performed against databases of different sizes. The molecules for docking come from the newly established COCONUT-like molecular dataset. We created four levels of virtual screening molecular libraries of different sizes by selecting the top 0.1%, 1.0%, 10.0%, and 100.0% of molecules from the natural product–like database. After that, we performed molecular docking with the previously identified conserved coronavirus proteins: SARS-CoV-2 3CLpro, RdRp, and nsp13.[36] All molecular docking tasks were performed using the Glide HTVS docking program in Schrödinger Release 2020.

RESULTS AND DISCUSSION

Molecular classification screening and statistics results for different datasets

Molecular classification statistics are performed on three datasets extracted from ChEMBL30, DrugBank (5.1.9), and COCONUT (January 2022), resulting in the final datasets containing 1,038,551, 11,283, and 406,920 molecules, respectively, after a simple filter.

Using D3Rings, acyclic, monocyclic, and joined ring compounds in the three datasets were screened (Figure 4 left and Table S1). The statistical results show substantial variations: 1) The proportion of molecules without any cyclic structure (blue) is higher in the DrugBank (13.77%) than that in the bioactive compound database ChEMBL (1.08%) and natural product database COCONUT (5.14%); 2) the ChEMBL and DrugBank datasets shared a similar proportion of molecules with one ring structure or no direct linkage between multiple rings (purple) (42.85% vs. 42.15%, respectively), contrasting with the lower content of such compounds in the natural product and natural product–like molecular datasets (23.28%); 3) COCONUT dataset had a substantially higher proportion of molecules with two or more directly linked rings (green) (71.58%) compared to ChEMBL and DrugBank (56.07% vs. 44.08%, respectively). In summary, ring compounds are the most abundant compounds in all three datasets, with ChEMBL having the fewest acyclic compounds, COCONUT having the most cyclic compounds, and DrugBank showcasing the simplest ring structures (the content of acyclic compounds and monocyclic compounds exceeds 55%). As expected, COCONUT has the most complicated structures (71.58% cyclic compounds; among the cyclic compounds, the proportion of joined ring compounds is more than 3/4).

The three-dimensional stacked bar chart on the right side of Figure 4 and Table S2 illustrates a detailed analysis of the joined ring compounds in three different ways: the presence of spiro ring structures (Classification 1), the presence of fused or bridged ring structures (Classification 2), and the maximum number of joined rings in fused ring compounds (Classification 3). Key findings include the following: 1) COCONUT database is richer in spiro ring compounds (5.38%) compared to other databases; 2) fused and bridged ring compounds dominate all three datasets, constituting 55.10%, 43.69%, and 71.15% in ChEMBL, DrugBank, and COCONUT, respectively; 3) COCONUT has the most structures with fused rings (the number of fused rings ≥ 2), especially, significantly more molecules with up to three or more fused ring structures. In addition, COCONUT contains more cage compounds than drug molecules and bioactive molecules, including some alkaloids, terpenoids, xanthones, and some special cage like structures extracted from marine organisms. The COCONUT database also contains more macrocyclic compounds than ChEMBL and DrugBank, accounting for approximately 3.6 and 2.3 times the proportion of macrocyclic compounds in ChEMBL and DrugBank databases, respectively (Figure 5, Table S1).

From these results, we found that the chemical structures of COCONUT compounds are more complex than that in ChEMBL and DrugBank, and are rich in characteristic structures such as spiro ring, fused rings, bridged rings and macrocycles.

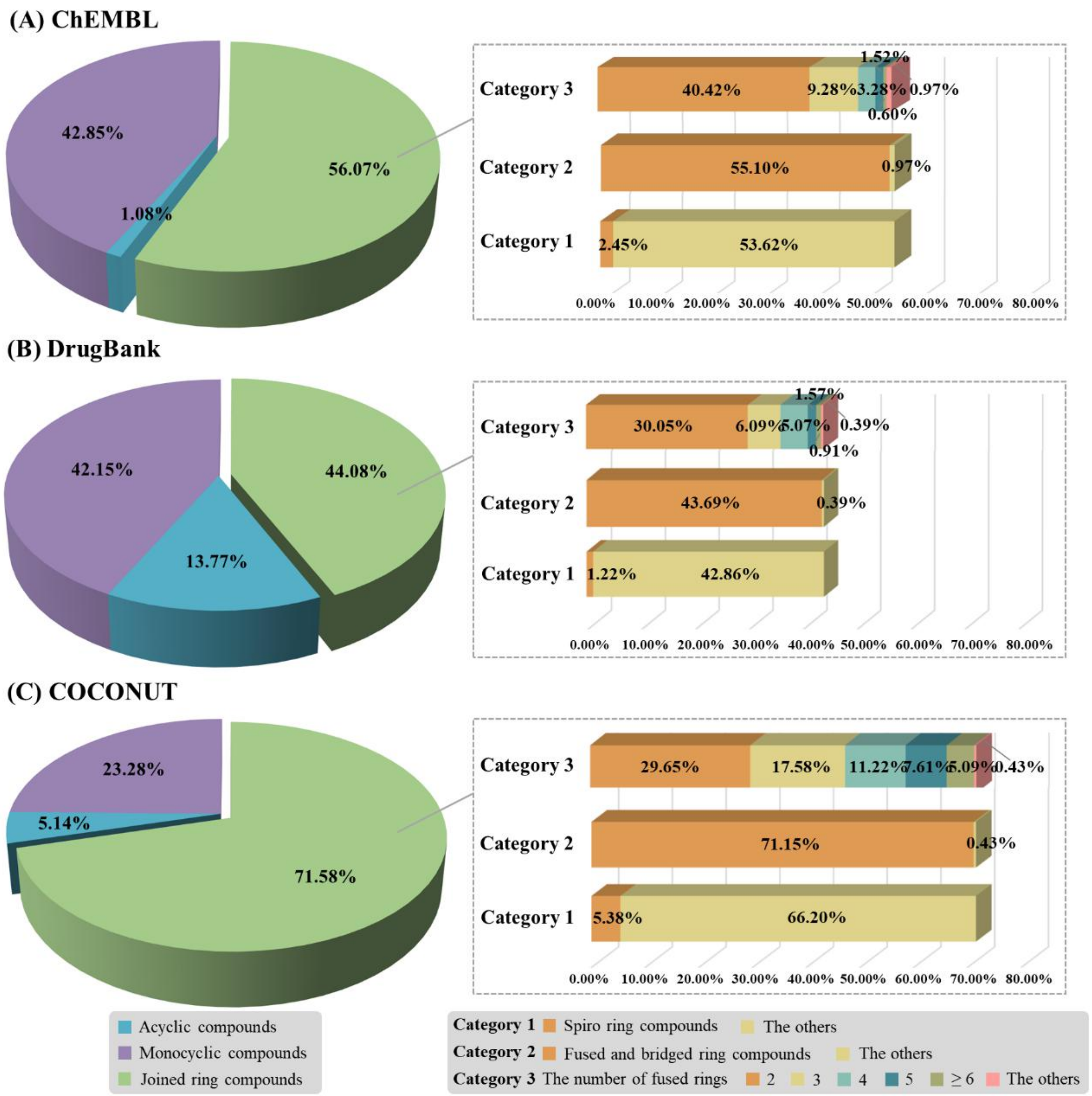

Figure 4. Molecular classification statistics for (A) ChEMBL30, (B) DrugBank (5.1.9), (C) COCONUT (January 2022) datasets.

* The proportions of acyclic, monocyclic, and joined ring compounds in the dataset are shown in the 3D pie chart on the left. The statistical results for the continued classification of joined ring compounds according to the presence of spiro ring structures (Category 1), the presence of fused and bridged ring structures (Category 2), and the number of fused rings contained (Category 3) are shown in the 3D stacked bar chart on the right.

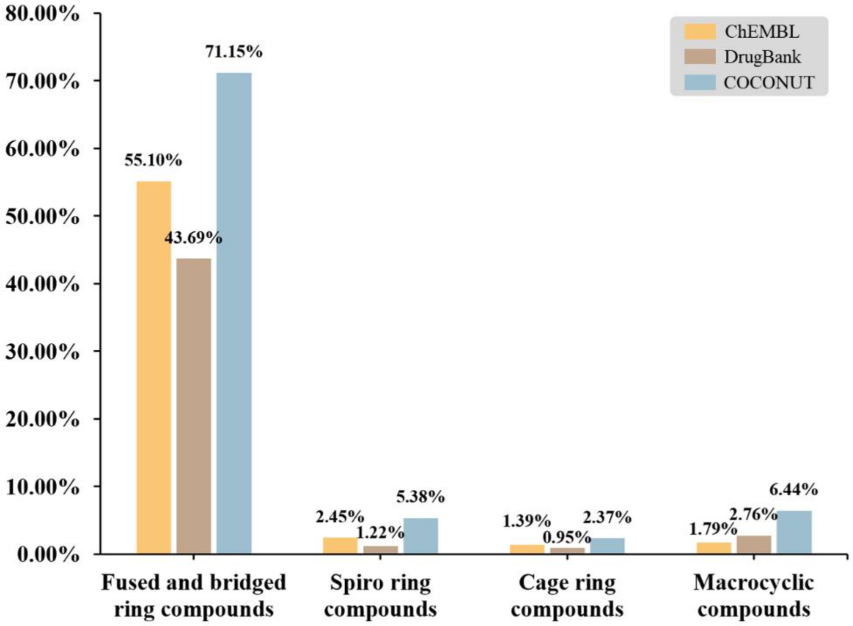

Figure 5. Proportions of the fused and bridged ring, spiro ring, cage ring, and macrocyclic compounds in ChEMBL30, DrugBank (5.1.9), and COCONUT (January 2022) datasets.

Statistical results for halogenated compounds in different molecular datasets

In our analysis of halogenated compounds in the ChEMBL, DrugBank, and COCONUT databases, a consistent trend emerged—bioactive compounds exhibited the highest proportion of halogenated compounds, followed by the DrugBank database, while natural or natural compound-like products showed the lowest proportion. This pattern was consistent across all sample pools, including spiro ring, fused and bridged ring, and macrocyclic compounds (Figures 6 and S1 and Tables S3 and S4). Despite the noteworthy presence of halogenated compounds among the drug-like compounds, the actual number of halogenated drugs reaching the market is smaller than expected. One reason for this is that natural products are an inspiration for discovering new drugs, but as previous research and our results show, halogenated natural products are rare.[37–39] From previous research, there were only 10,310 halogenated natural products described in the Dictionary of Natural Products (DNP) in March 2021, while we found 40,722 halogenated natural products and halogenated natural product–like molecules in the COCONUT database, a substantial increase compared to the 10,310 halogenated natural products documented in the Dictionary of Natural Products as of March 2021.[40,41] Figures 6 and S1 also illustrate that fluorine- and chlorine-containing compounds are the most common and bromine- and iodine-containing compounds are scarce. In fact, fluorine- and chlorine-containing compounds are more stable than iodine-containing compounds. In addition, as shown in Figure S2 and Table S5, most halogenated compounds, especially in halogenated natural products, contain only one halogen atom (up to 61% of halogenated natural products contain only one halogen atom). In contrast, halogenated drugs contain a higher proportion of compounds with multiple halogen atoms (nearly 50%), highlighting the importance of introducing halogen atoms in modulating the lipophilicity, pKa, conformation, and bioavailability of drugs.[42]

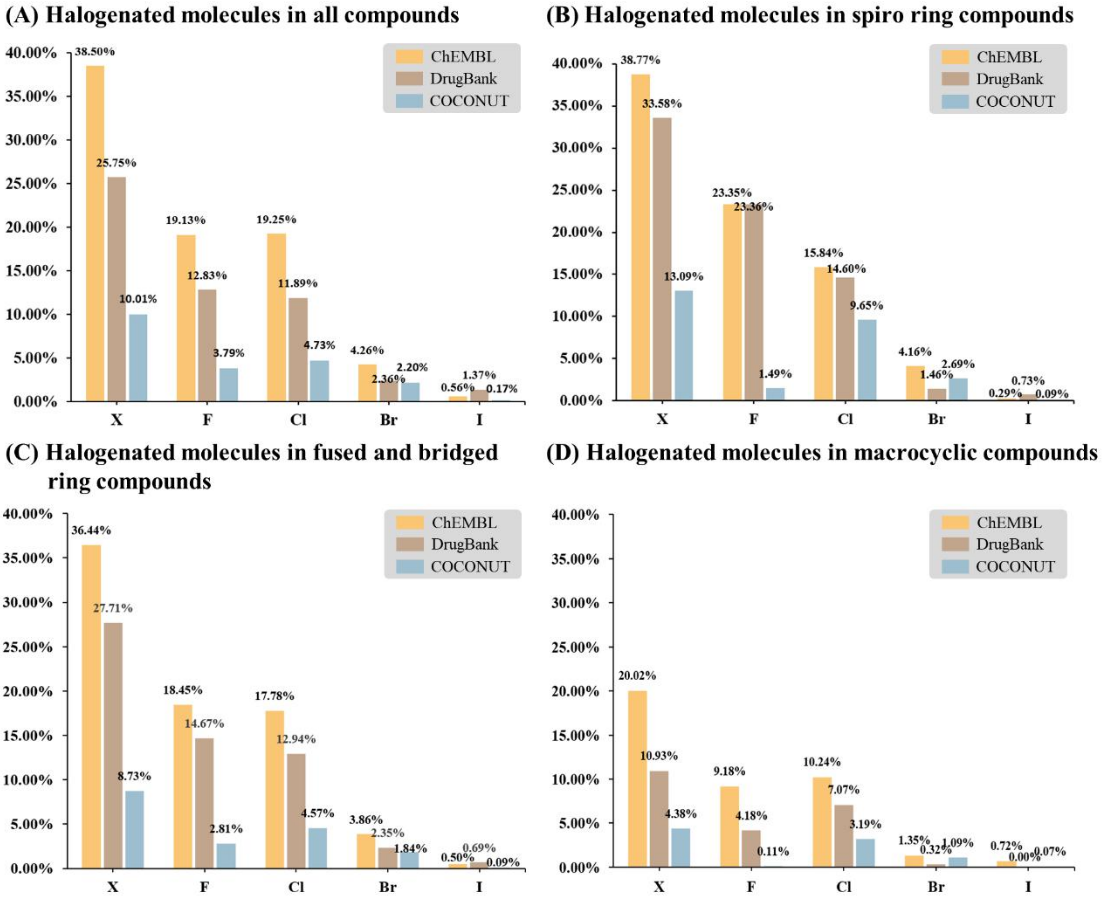

Figure 6. Proportions of halogen-containing compounds in ChEMBL30, DrugBank (5.1.9), and COCONUT (January 2022) datasets.

Generation and establishment of large molecular datasets

The VAE, AAE, and CharRNN models were trained using the DrugBank or COCONUT training sets, respectively, with the corresponding training hyperparameters obtained after parameter tuning shown in Table S6.

The VAE, AAE, and CharRNN models were trained using a set of 10,154 molecules from DrugBank (5.1.9), generating 40,000 valid molecules each. After merging and deduplicating, a drug-like molecule dataset of 119,381 molecules was established, approximately 10 times larger than the DrugBank dataset. Similarly, the VAE, AAE, and CharRNN models were trained using a set of 366,228 molecules from COCONUT (January 2022), generating 1,500,000 valid molecules each. After merging and deduplicating, a natural product–like molecular dataset of 4,185,929 molecules was obtained, approximately 10 times the size of the COCONUT dataset. To assess the generated molecules, we randomly selected 10,000 molecules from the training set and each of the three generated molecule sets and downscaled them using the t-SNE dimensionality reduction algorithm, and the results are shown in Figure 7. The distribution in reduced dimensions shows substantial overlap between the generated molecules and their respective training data from DrugBank or COCONUT, indicating that the generated molecules reproduce the properties of the molecules in the training set very well. Furthermore, the different molecular deep generative models complement each other and explore chemical space beyond the molecules in the training set. Figure S3 showcases randomly selected molecules from the DrugBank- and COCONUT-like datasets generated using VAE, AAE, and CharRNN. The structures of the newly generated molecules are valid and chemically reasonable.

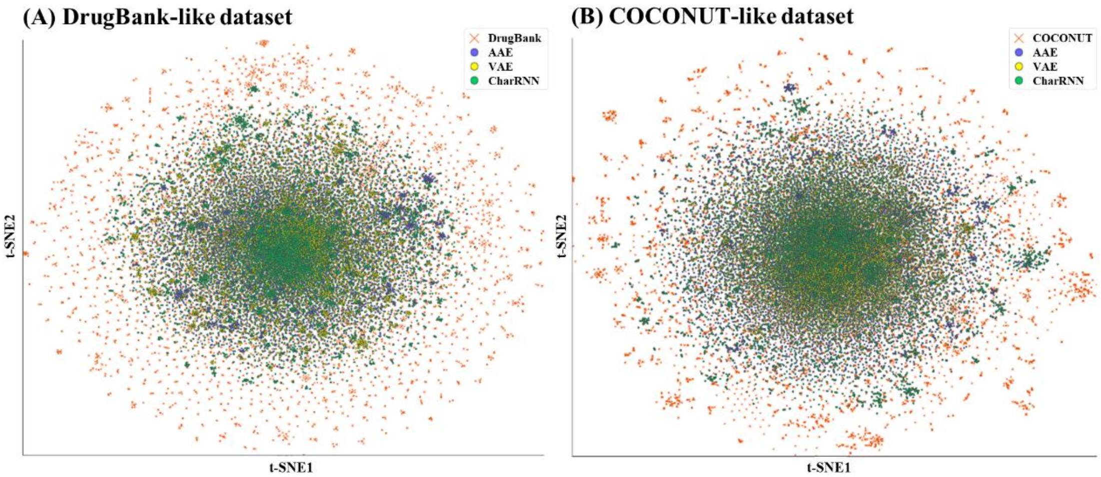

Figure 7. t-SNE dimensionality reduction plot of sampled molecules in training sets and generated molecule sets for (A) DrugBank-like dataset and (B) COCONUT-like dataset.

Evaluation of molecular deep generative models and statistics for generated molecular databases

The models were trained on the DrugBank or COCONUT training sets, and their performance in generating molecules was evaluated, with results shown in Table 1. All models performed well in evaluating the molecular validity, novelty, internal diversity, and uniqueness of the top 10,000 valid molecules generated using the models (Unique 10k). Validity indicators of the generated molecules approached or equaled 1, indicating the validity of SMILES strings for all generated molecules. The proportion of generated molecules that did not exist in the training set (Novelty) exceeded 99%, suggesting that none of the models had overfitting problems, and the internal diversity of the generated molecules was evaluated with scores ranging from 0.888 to 0.902, showing that these models were beneficial for discovering novel chemical structures and expanding chemical space. Uniqueness evaluation scores were all greater than 0.98, demonstrating that these models were not limited to producing only a few types of molecules. In the similarity evaluation of compound substructure distribution, the BRICS fragments (Frag) of generated molecules showed higher similarity to the training set, while the Bemis–Murcko molecular scaffolds (Scaff) generated through the three models trained on COCONUT showed considerably lower similarity to the training set compared to models trained on DrugBank. This suggests the emergence of more novel molecular scaffolds in the COCONUT-like dataset. Finally, the assessed values of SNN for each model remained between 0.3–0.5, combined with the high novelty results of the model, indicating that the models did not suffer from overfitting and explored new chemical space.

Table 1. Performance metrics for models after training with DrugBank or COCONUT training sets.

	Valid (↑)	Novelty (↑)	Internal diversity (↑)	Unique 10k (↑)	Frag (↑)	Scaff (↑)	SNN (↑)
			DrugBank-like dataset				
AAE	1.0	0.999	0.892	1.0	0.977	0.834	0.451
VAE	1.0	0.999	0.888	1.0	0.998	0.907	0.486
CharRNN	1.0	0.999	0.888	0.999	0.998	0.931	0.556
			COCONUT-like dataset				
AAE	0.999	0.998	0.902	0.991	0.979	0.461	0.334
VAE	1.0	0.998	0.899	0.986	0.969	0.307	0.331
CharRNN	0.999	0.990	0.898	0.993	0.993	0.524	0.373

* The metrics include molecular validity (Valid), molecular novelty (Novelty), internal diversity (Internal diversity), uniqueness (Unique 10k), fragment similarity (Frag), scaffold similarity (Scaff), and similarity to the nearest neighbor (SNN). The limits of these metrics are [0,1], where larger values indicate better performance.

In addition, we compared the distributions of MW, LogP, QED, and SA scores for compounds in the training set against those generated through DrugBank- or COCONUT-like models (Figure 8). Figure 8A shows that the distributions of all chemical properties for the molecules generated through deep generative models overlap well with those of DrugBank molecules (orange), indicating effective capturing of the molecular characteristics of the training set. Figure 8B shows that the models can reproduce the properties of the training set molecules well, with the QED and SA distribution graphs indicating that the properties of the molecules generated through the CharRNN model closely overlap with those of the training set (orange).

Additionally, we evaluated the performance of AAE, VAE, and CharRNN models and compared them with LatentGAN (Prykhodko et al., 2019) and JT-VAE (Jin et al., 2018) models using the MOSES benchmarking score metrics.[43,44] All models were trained on the DrugBank dataset, and 10,000 molecules each were subsequently generated for evaluation. Table S7 reveals that LatentGAN and JT-VAE models perform substantially worse compared to the other three models in internal diversity, Unique 10k, Frag, Scaff, and SNN (values shown in gray shading). Figure S4 shows the distributions of four molecular properties in the generated and test sets. JT-VAE is biased toward lighter molecules, indicating ease of molecule synthesis (SA) but low drug-like properties (QED). LatentGAN prefers to generate molecules with excessive MW, consistent with its low QED and high SA score. In contrast, the other three models closely match the data distribution.

The statistical results for acyclic, monocyclic, and joined ring compounds in the DrugBank- and COCONUT-like databases are shown in the 3D pie chart on the left column of Figure S5 and Table S8. The joined ring compounds are further divided into spiro ring compounds, fused and bridged ring compounds, and the number of fused rings contained, as shown in the 3D stacked bar chart on the right column of Figure S5 and Tables S8 and S9. Additionally, Figure S6 compares the percentages of spiro ring, fused and bridged ring, cage ring, and macrocyclic compounds in different generated molecular datasets. The statistical results for halogenated compounds are shown in Figures S7–S9 and Tables S10–S12. These statistical results show that the prevalence of each type of feature molecule in the DrugBank- and COCONUT-like datasets aligns with that in the original DrugBank and COCONUT databases. Notably, the proportion of molecules with a maximum fused ring size of three or more (Category 3 in Figure S5) in the generated molecular dataset is smaller than that in the original dataset. This suggests that under the condition of limited training data on structurally complex cyclic molecules, deep generative models exhibit a reduced likelihood of generating molecules with more complex ring structures.

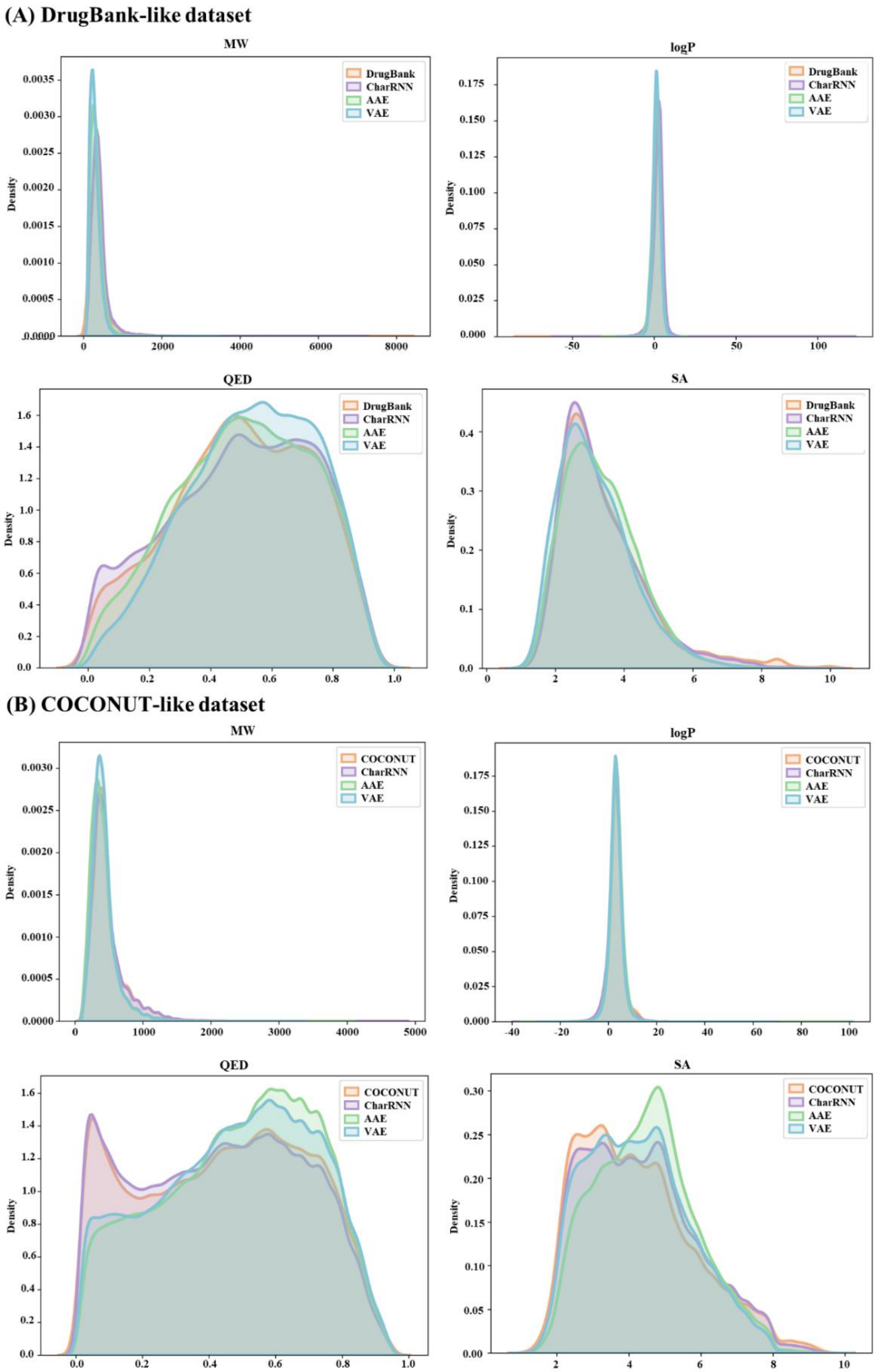

Figure 8. Distribution of chemical properties for (A) DrugBank training set and the DrugBank-like dataset generated through VAE, AAE, and CharRNN and (B)

COCONUT training set and the COCONUT-like dataset generated through VAE, AAE, and CharRNN.

* Key metrics: (1) MW: molecular weight. (2) LogP: oil–water partition coefficient. (3) QED: quantitative estimation of drug-likeness, a [0,1] value estimating the potential of a molecule as a drug candidate. (4) SA: synthetic accessibility score, a heuristic estimate of synthesis difficulty of given molecules: hard (10) or easy (1).

The impact of molecular dataset size on virtual screening

As the size of the virtual screening library expands, the ability to find molecules with high target affinity influences the optimal size of the molecular database. A total of 4,185,929 molecules were included in the COCONUT-like molecular dataset, from which subsets of 4,186 (0.1%), 41,859 (1%), 418,593 (10%), and 4,185,929 (100%) molecules were sequentially extracted to construct four subsets of different sizes and docked against the SARS-CoV-2 conserved proteins 3CLpro, RdRp, and nsp13. Figure 9A illustrates that the number of molecules with a high likelihood of binding to receptor proteins increases with the expansion of the docking library size (Table S13). As the virtual screening library size increases by a factor of 1,000, the number of molecules with docking scores less than −6.0 kcal/mol against the target proteins also experiences an approximately 1,000-fold increase.

As depicted in Figure 9B and Table S13, the enhancement of best docking scores correlates with the expansion of the molecular library. The best docking score for SARS-CoV-2 3CLpro increased from −8.408 to −10.693 kcal/mol (Δ_{3CLpro} = −2.285 kcal/mol), for SARS-CoV-2 RdRp from −10.148 to −12.112 kcal/mol (Δ_{RdRp} = −1.964 kcal/mol), and for SARS-CoV-2 nsp13 from −10.944 to −11.864 kcal/mol (Δ_{nsp13} = −0.920 kcal/mol). The improvement in best docking scores was monotonic with the expansion of the docking library. As the docking subset increased by a factor of 10^3, the mean scores of the top 100 docked molecules monotonically increased in virtual screening against SARS-CoV-2 3CLpro, RdRp, and nsp13, with Δ_{3CLpro} = −2.246 kcal/mol, Δ_{RdRp} = −2.440 kcal/mol, and Δ_{nsp13} = −2.227 kcal/mol, respectively (Figure 9C). This suggests that as the docking molecular library expands, the predicted docking scores of top-ranked molecules to the targets steadily improve, and showing no signs of saturation in our experiment.

In addition, the compounds involved in docking were categorized into five subsets: acyclic, monocyclic, spiro ring, fused and bridged ring, and macrocyclic compounds. Figure S10A and B shows that the best docking scores are always observed from the molecules with ring structures, with monocyclic compounds and fused and bridged ring compounds performing better in the docking results against the three targets. Figure S10C–D shows that monocyclic compounds performed best in docking with RdRp targets, while fused and bridged ring compounds showed the best affinity against 3CLpro and nsp13.

Therefore, virtual screening of a large library of molecules proves beneficial for improving docking results and finding ligands with good affinity to the targets. This observation aligns with a study by Lyu et al., where ultralarge libraries were docked against D4, σ2, and 5HT2A receptors to study how docking scores varied with library size. The results indicated that as the library grew from 10^5 to over 10^9 molecules, the fit of the top-ranking docked molecules steadily improved without reaching saturation and the number of molecules in the favorable scoring region increased.[35] In our experiments, cyclic structures, such as fused and bridged ring and monocyclic compounds, were prevalent and exhibited the best docking scores among the top-ranked docking molecules.

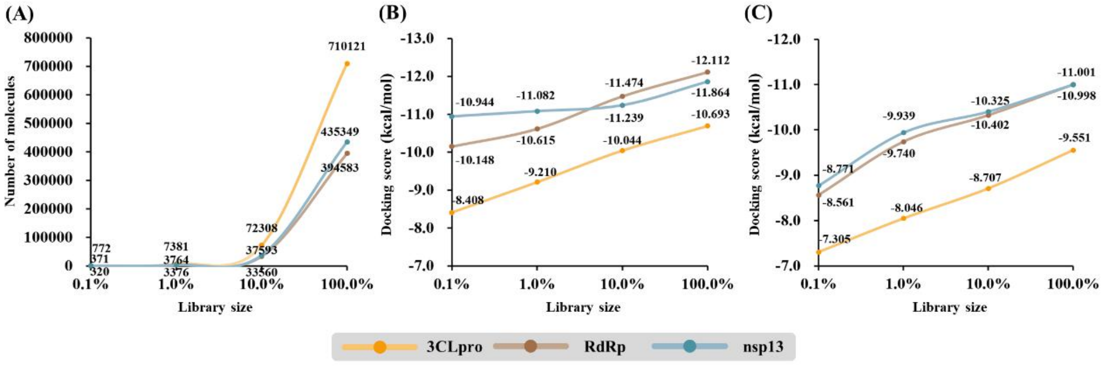

Figure 9. Effect of library size on docking performance against the SARS-CoV-2 3CLpro, RdRp, and nsp13.

* (A) Change in the number of molecules with docking scores less than −6.0 kcal/mol with

increasing library size. (B) Change of the best docking scores as library size grows. (C) Change in the average docking scores of the top-ranking 100 molecules with increasing library size.

CONCLUSION

Cyclic compounds are ubiquitous in nature, making them an important category of molecules for drug discovery and development. Herein, we developed a new method, called D3Rings, that can be used to identify acyclic, monocyclic, spiro ring, fused and bridged ring, and macrocyclic compounds. With this method, we performed a statistical analysis of cyclic compounds in three different molecular datasets (ChEMBL, DrugBank, and COCONUT) and found that the natural product–like database COCONUT is rich in cyclic structures, such as spiro ring, fused and bridged rings, cage ring compounds, and macrocycles. Leveraging three well-trained deep generative models (e.g., VAE, AAE, and CharRNN), we generated ten times larger drug-like and natural product–like molecular datasets than DrugBank and COCONUT, respectively. Docking the newly generated COCONUT-like database to three anti-COVID-19 target proteins reveals that as the molecular library expanded, the docked binding affinity between the top-ranked docked molecules and target proteins steadily improved. Our findings underscore the practical value of larger molecular datasets containing cyclic compounds for future drug discovery. This work underscores the importance of cyclic compounds and the potential impact of enriched molecular datasets in advancing drug discovery and development.

ASSOCIATED CONTENT

Supporting information

The Supporting Information is available free of charge on the ACS Publications website at DOI: XXX.

The proportions of halogenated compounds in ChEMBL30, DrugBank, and COCONUT datasets (Figures S1 and S2). Randomly selected molecules generated through VAE, AAE, and CharRNN (Figure S3). Distribution of chemical properties for DrugBank training set versus the molecular dataset generated through VAE, AAE, CharRNN, LatentGAN, and JT-VAE (Figure S4). Molecular classification statistics for DrugBank-like and COCONUT-like datasets (Figures S5 and S6). The proportions of halogenated compounds in DrugBank-like and COCONUT-like datasets (Figures S7–S9). The performance of different structurally characterized compounds docking against receptors (Figure S10).

Molecular classification statistics for ChEMBL, DrugBank, and COCONUT datasets (Tables S1 and S2). Number and proportions of halogenated compounds in ChEMBL, DrugBank, and COCONUT datasets (Tables S3–S5). Hyperparameters of VAE, AAE, and CharRNN (Table S6). Performance metrics for AAE, VAE, CharRNN, LatentGAN, and JT-VAE models (Table S7). Molecular classification statistics for DrugBank-like and COCONUT-like datasets (Tables S8–S9). Number and proportions of halogenated compounds in DrugBank-like and COCONUT-like datasets (Tables S10–S12). The performance of docking against SARS-CoV-2 3CLpro, RdRp, and nsp13 receptors changes with the expansion of molecular library size (Table S13).

AUTHOR INFORMATION

Corresponding authors

Zhijian Xu−State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; orcid.org/0000-0002-3063-8473; Email: zjxu@simm.ac.cn.

Weiliang Zhu−State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China; orcid.org/0000-0001-6699-5299; Phone: +86-21-50805020; Email: wlzhu@simm.ac.cn.

Authors

Minfei Ma–State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

Xinben Zhang–State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica,Chinese Academy of Sciences, Shanghai 201203, China.

Liping Zhou–State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

Zijian Han–State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

Yulong Shi–State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

Jintian Li–State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

Leyun Wu–State Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

Author contributions

Zhijian Xu and Weiliang Zhu designed and supervised the study. Minfei Ma performed the study. Xinben Zhang built the website. Liping Zhou, Zijian Han, Yulong Shi, Jintian Li, Leyun Wu helped to analyze data. Minfei Ma, Weiliang Zhu, and Zhijian Xu drafted the initial manuscript.

Conflict of interest

The authors declare no competing financial interest.

Data and software availability

The SMILES string representations of the molecules used in this study were obtained from the ChEMBL (https://www.ebi.ac.uk/chembl/), DrugBank (https://go.drugbank.com/) and COCONUT (https://coconut.naturalproducts.net/)

databases, respectively. Molecular Sets (MOSES), the benchmark platform for molecular generation models used in this study was openly available at https://github.com/molecularsets/moses. The software package of Schrödinger maestro 10.2.010 (https://www.schrodinger.com/downloads/releases) was provided by Schrödinger.

ACKNOWLEDGMENTS

This study was supported from the National Key Research and Development Program of China (2022YFA1004304) and the National Natural Science Foundation of China (82273851 and 82322067).

REFERENCES

(1) Owen, D. R.; Allerton, C. M. N.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; Dantonio, A.; Di, L.; Eng, H.; Ferre, R.; Gajiwala, K. S.; Gibson, S. A.; Greasley, S. E.; Hurst, B. L.; Kadar, E. P.; Kalgutkar, A. S.; Lee, J. C.; Lee, J.; Liu, W.; Mason, S. W.; Noell, S.; Novak, J. J.; Obach, R. S.; Ogilvie, K.; Patel, N. C.; Pettersson, M.; Rai, D. K.; Reese, M. R.; Sammons, M. F.; Sathish, J. G.; Singh, R. S. P.; Steppan, C. M.; Stewart, A. E.; Tuttle, J. B.; Updyke, L.; Verhoest, P. R.; Wei, L.; Yang, Q.; Zhu, Y. An Oral SARS-CoV-2 Mpro Inhibitor Clinical Candidate for the Treatment of COVID-19. *Science* **2021**, *374* (6575), 1586–1593. DOI: 10.1126/science.abl4784.

(2) Tyndall, J. D. A. S-217622, a 3CL Protease Inhibitor and Clinical Candidate for SARS-CoV-2. *J. Med. Chem.* **2022**, *65* (9), 6496–6498. DOI: 10.1021/acs.jmedchem.2c00624.

(3) Sheahan, T. P.; Sims, A. C.; Zhou, S.; Graham, R. L.; Pruijssers, A. J.; Agostini, M. L.; Leist, S. R.; Schäfer, A.; Dinnon, K. H.; Stevens, L. J.; Chappell, J. D.; Lu, X.; Hughes, T. M.; George, A. S.; Hill, C. S.; Montgomery, S. A.; Brown, A. J.; Bluemling, G. R.; Natchus, M. G.; Saindane, M.; Kolykhalov, A. A.; Painter, G.; Harcourt, J.; Tamin, A.; Thornburg, N. J.; Swanstrom, R.; Denison, M. R.; Baric, R. S.

An Orally Bioavailable Broad-Spectrum Antiviral Inhibits SARS-CoV-2 in Human Airway Epithelial Cell Cultures and Multiple Coronaviruses in Mice. *Sci. Transl. Med.* **2020**, *12* (541), eabb5883. DOI: 10.1126/scitranslmed.abb5883.

(4) Thompson, M. G.; Stenehjem, E.; Grannis, S.; Ball, S. W.; Naleway, A. L.; Ong, T. C.; DeSilva, M. B.; Natarajan, K.; Bozio, C. H.; Lewis, N.; Dascomb, K.; Dixon, B. E.; Birch, R. J.; Irving, S. A.; Rao, S.; Kharbanda, E.; Han, J.; Reynolds, S.; Goddard, K.; Grisel, N.; Fadel, W. F.; Levy, M. E.; Ferdinands, J.; Fireman, B.; Arndorfer, J.; Valvi, N. R.; Rowley, E. A.; Patel, P.; Zerbo, O.; Griggs, E. P.; Porter, R. M.; Demarco, M.; Blanton, L.; Steffens, A.; Zhuang, Y.; Olson, N.; Barron, M.; Shifflett, P.; Schrag, S. J.; Verani, J. R.; Fry, A.; Gaglani, M.; Azziz-Baumgartner, E.; Klein, N. P. Effectiveness of Covid-19 Vaccines in Ambulatory and Inpatient Care Settings. *N. Engl. J. Med.* **2021**, *385* (15), 1355–1371. DOI: 10.1056/NEJMoa2110362.

(5) Qian, H. J.; Wang, Y.; Zhang, M. Q.; Xie, Y. C.; Wu, Q. Q.; Liang, L. Y.; Cao, Y.; Duan, H. Q.; Tian, G. H.; Ma, J.; Zhang, Z. B.; Li, N.; Jia, J. Y.; Zhang, J.; Aisa, H. A.; Shen, J. S.; Yu, C.; Jiang, H. L.; Zhang, W. H.; Wang, Z.; Liu, G. Y. Safety, Tolerability, and Pharmacokinetics of VV116, an Oral Nucleoside Analog Against SARS-CoV-2, in Chinese Healthy Subjects. *Acta Pharmacol. Sin.* **2022**, *43* (12), 3130–3138. DOI: 10.1038/s41401-022-00895-6.

(6) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. DOI: 10.1021/jm9602928.

(7) Shang, J.; Sun, H.; Liu, H.; Chen, F.; Tian, S.; Pan, P.; Li, D.; Kong, D.; Hou, T. Comparative Analyses of Structural Features and Scaffold Diversity for Purchasable Compound Libraries. *J. Cheminform.* **2017**, *9* (1), 25. DOI: 10.1186/s13321-017-0212-4.

(8) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree-Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47* (1), 47–58. DOI: 10.1021/ci600338x.

(9) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting Biologically Relevant Chemical Space: A Structural Classification of Natural Products (SCONP). *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*

(48), 17272–17277. DOI: 10.1073/pnas.0503647102.

(10) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Review of Ring Perception Algorithms for Chemical Graphs. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (3), 172–187. DOI: 10.1021/ci00063a007.

(11) Lee, C. J.; Kang, Y. M.; Cho, K. H.; A Robust Method for Searching the Smallest Set of Smallest Rings with A Path-Included Distance Matrix. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (41), 17355–17358. DOI: 10.1073/pnas.0813040106.

(12) Lavecchia, A. Deep Learning in Drug Discovery: Opportunities, Challenges and Future Prospects. *Drug Discov. Today* **2019**, *24* (10), 2017–2032. DOI: 10.1016/j.drudis.2019.07.006.

(13) Öztürk, H.; Özgür, A.; Schwaller, P.; Laino, T.; Ozkirimli, E. Exploring Chemical Space Using Natural Language Processing Methodologies for Drug Discovery. *Drug Discov. Today* **2020**, *25* (4), 689–705. DOI: 10.1016/j.drudis.2020.01.020.

(14) Skalic, M.; Sabbadin, D.; Sattarov, B.; Sciabola, S.; De Fabritiis, G. From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design. *Mol. Pharm.* **2019**, *16* (10), 4282–4291. DOI: 10.1021/acs.molpharmaceut.9b00634.

(15) Vogt, M. Exploring Chemical Space — Generative Models and Their Evaluation. *Artif. Intell. Life Sci.* **2023**, *3*, 100064. DOI: 10.1016/j.ailsci.2023.100064.

(16) Bian, Y.; Xie, X. Q. Generative Chemistry: Drug Discovery with Deep Learning Generative Models. *J. Mol. Model.* **2021**, *27* (3), 71. DOI: 10.1007/s00894-021-04674-8.

(17) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. DOI: 10.1093/nar/gky1075.

(18) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson,

M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D1074–D1082. DOI: 10.1093/nar/gkx1037.

(19) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminform.* **2021**, *13* (1), 2. DOI: 10.1186/s13321-020-00478-9.

(20) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. DOI: 10.1021/ci00057a005.

(21) Zheng, Y.; Tice, C. M.; Singh, S. B. The Use of Spirocyclic Scaffolds in Drug Discovery. *Bioorg. Med. Chem. Lett.* **2014**, *24* (16), 3673–3682. DOI: 10.1016/j.bmcl.2014.06.081.

(22) Li, Y.; Zhang, L.; Wang, W.; Liu, Y.; Sun, D.; Li, H.; Chen, L. A Review on Natural Products with Cage-Like Structure. *Bioorg. Chem.* **2022**, *128*, 106106. DOI: 10.1016/j.bioorg.2022.106106.

(23) Bai, H.; Wang, J.; Li, Z.; Tang, G. Macrocyclic Compounds for Drug and Gene Delivery in Immune-Modulating Therapy. *Int. J. Mol. Sci.* **2019**, *20* (9), 2097. DOI: 10.3390/ijms20092097.

(24) Kingma, D. P.; Welling, M. Auto-encoding Variational Bayes. arXiv December 10, **2022**. DOI: 10.48550/arXiv.1312.6114.

(25) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112* (518), 859–877. DOI: 10.1080/01621459.2017.1285773.

(26) Commenges, D. Information Theory and Statistics: An Overview. arXiv November 3, **2015**. DOI: 10.48550/arXiv.1511.00860.

(27) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for de Novo Molecular Design. *J. Cheminform.* **2018**, *10* (1), 31. DOI: 10.1186/s13321-018-0286-7.

(28) Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial Autoencoders. arXiv May 24, **2016**. DOI: 10.48550/arXiv.1511.05644.

(29) Hong, S. H.; Ryu, S.; Lim, J.; Kim, W. Y. Molecular Generative Model Based on an

Adversarially Regularized Autoencoder. *J. Chem. Inf. Model.* **2020**, *60* (1), 29–36. DOI: 10.1021/acs.jcim.9b00694.

(30) Bjerrum, E. J.; Threlfall, R. Molecular Generation with Recurrent Neural Networks (RNNs). arXiv May 17, **2017**. DOI: 10.48550/arXiv.1705.04612.

(31) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (3), 1175–1183. DOI: 10.1021/acs.jcim.9b00943.

(32) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 565644. DOI: 10.3389/fphar.2020.565644.

(33) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. DOI: 10.1038/s41586-019-0917-9.

(34) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X. P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; Zvonok, N.; Jain, M. K.; Savych, O.; Radchenko, D. S.; Nikas, S. P.; Petasis, N. A.; Moroz, Y. S.; Roth, B. L.; Makriyannis, A.; Katritch, V. Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds. *Nature* **2022**, *601* (7893), 452–459. DOI: 10.1038/s41586-021-04220-9.

(35) Lyu, J.; Irwin, J. J.; Shoichet, B. K. Modeling the Expansion of Virtual Screening Libraries. *Nat. Chem. Biol.* **2023**, *19* (6), 712–718. DOI: 10.1038/s41589-022-01234-w.

(36) Ma, M.; Yang, Y.; Wu, L.; Zhou, L.; Shi, Y.; Han, J.; Xu, Z.; Zhu, W. Conserved Protein Targets for Developing Pan-Coronavirus Drugs Based on Sequence and 3D Structure Similarity Analyses. *Comput. Biol. Med.* **2022**, *145*, 105455. DOI: 10.1016/j.compbiomed.2022.105455: 0.1016.

(37) Gribble, G. W. Natural Organohalogens: A New Frontier for Medicinal Agents? *J.*

Chem. Educ. **2004**, *81* (10), 1441. DOI: 10.1021/ed081p1441.

(38) Lu, Y.; Liu, Y.; Xu, Z.; Li, H.; Liu, H.; Zhu, W. Halogen Bonding for Rational Drug Design and New Drug Discovery. *Expert Opin. Drug Discov.* **2012**, *7* (5), 375–383. DOI: 10.1517/17460441.2012.678829.

(39) Gribble, G. W. The Diversity of Naturally Produced Organohalogens. *Chemosphere* **2003**, *52* (2), 289–297. DOI: 10.1016/S0045-6535(03)00207-8.

(40) Sorokina, M.; Steinbeck, C. Review on Natural Products Databases: Where to Find Data in 2020. *J. Cheminform.* **2020**, *12* (1), 20. DOI: 10.1186/s13321-020-00424-9.

(41) Cochereau, B.; Meslet-Cladière, L.; Pouchus, Y. F.; Grovel, O.; Roullier, C. Halogenation in Fungi: What Do We Know and What Remains to Be Discovered? *Molecules* **2022**, *27* (10), 3157. DOI: 10.3390/molecules27103157.

(42) Shinada, N. K.; de Brevern, A. G.; Schmidtke, P. Halogens in Protein–Ligand Binding Mechanism: A Structural Perspective. *J. Med. Chem.* **2019**, *62* (21), 9341–9356. DOI: 10.1021/acs.jmedchem.8b01453.

(43) Prykhodko, O.; Johansson, S. V.; Kotsias, P. C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A De Novo Molecular Generation Method Using Latent Vector Based Generative Adversarial Network. *J. Cheminform.* **2019**, *11* (1), 74. DOI: 10.1186/s13321-019-0397-9.

(44) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation Artificial Intelligence in Drug Discovery, 2020, 228–249.

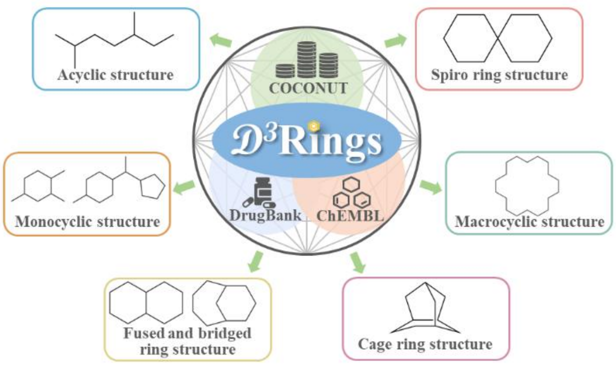

Table of Contents